Overview of the BlueGene/L System Architecture

Ray Bair

Director, Laboratory Computing Resource Center and TeraGrid Science Coordinator

Argonne National Laboratory, and The University of Chicago

3/1/06

THE UNIVERSITY OF CHICAGO

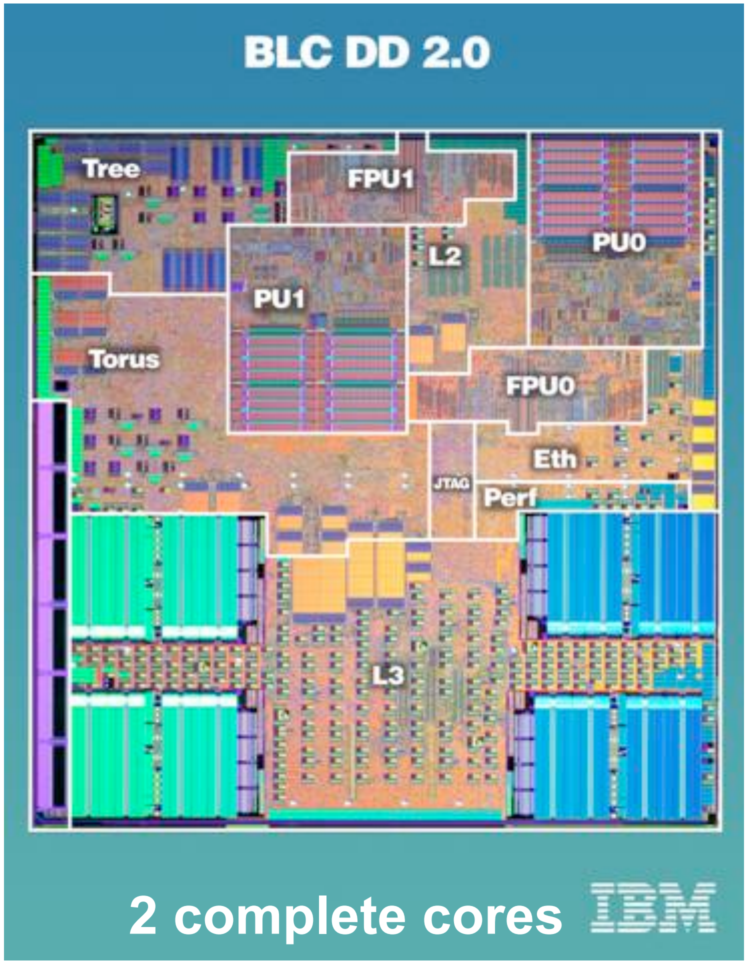
Office of Science
U.S. DEPARTMENT OF ENERGY

Argonne National Laboratory is managed by The University of Chicago for the U.S. Department of Energy

Overview of the BG/L System Architecture

- What's in the black box?
- How is it different than other computers?
- Will BlueGene like my application?

BlueGene/L Chip



Just add DRAM

Processor

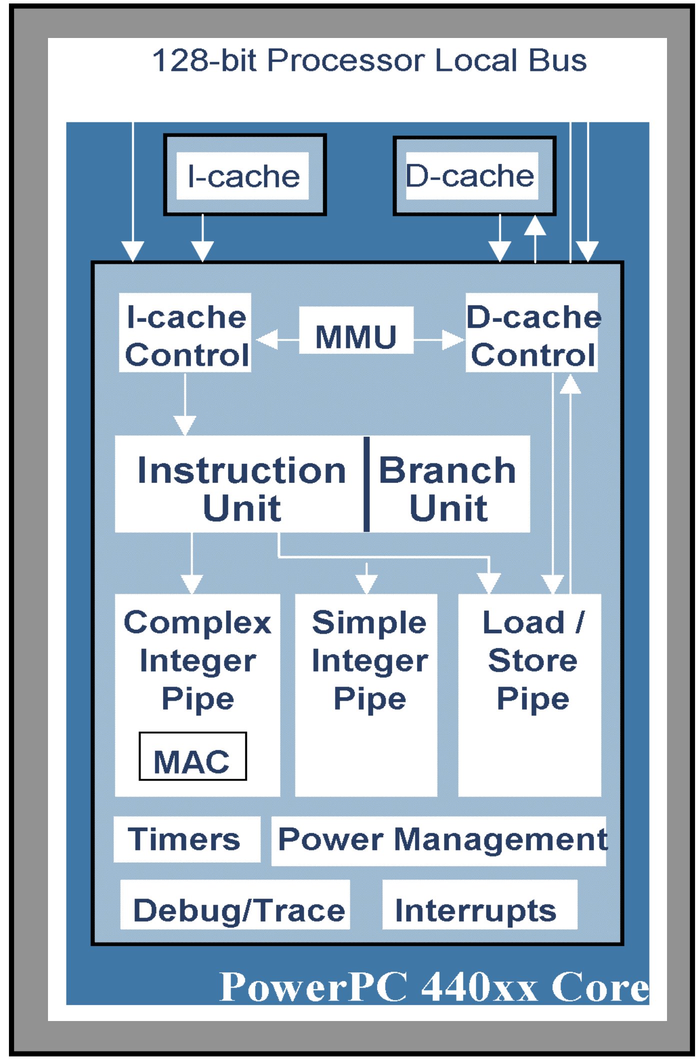
- PPC440x5 Processor Core – 700 MHz
  - *Superscalar: 2 instructions per cycle*
  - *Out of order issue and execution*
  - *Dynamic branch prediction, etc.*
- Two 64-bit floating point units
  - *SIMD instruct. over both register files*
  - *Parallel (quadword) loads/stores*
  - *2.8 GFLOPS/processor*

Interconnect

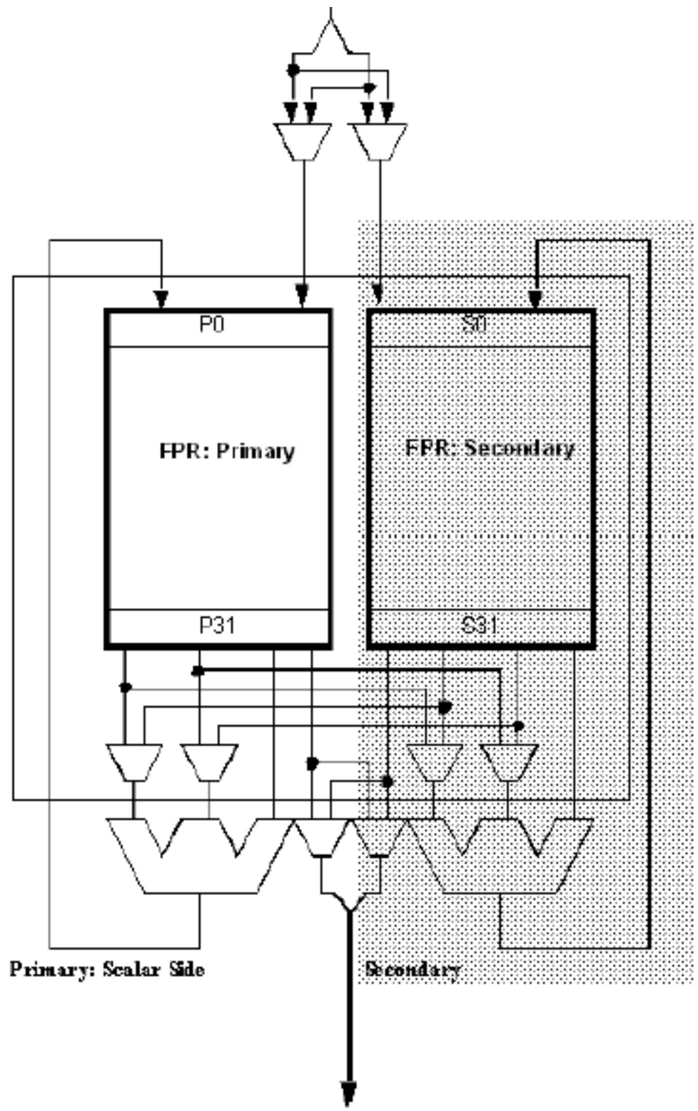
- 3 Dimensional Torus
  - *Virtual cut-through hardware routing*
  - *1.4Gb/s on all 12 node links*
  - *1 μs latency bet. neighbors, 5 μs to farthest*
- Global Tree
  - *One-to-all broadcast, reduction functionality*
  - *2.8 Gb/s of bandwidth per link*
  - *Latency of one way tree traversal 2.5 μs*
- Low Latency Global Barrier and Interrupt
  - *Latency of round trip 1.3 μs*
- Ethernet
  - *All external comm. (file I/O, control, etc.)*
- Control Network

PPC440x5 Processor Core Features

- High performance embedded PowerPC core
- 2.0 DMIPS/MHz
- Book E Architecture
- Superscalar: Two instructions per cycle
- Out of order issue, execution, and completion
- 7 stage pipeline
- 3 Execution pipelines
  - Combined complex, integer, & branch pipeline
  - Simple integer pipeline
  - Load/store pipeline.
- Dynamic branch prediction
- Single cycle multiply
- Single cycle multiply-accumulate
- Real-time non-invasive trace
- 128-bit CoreConnect Interface

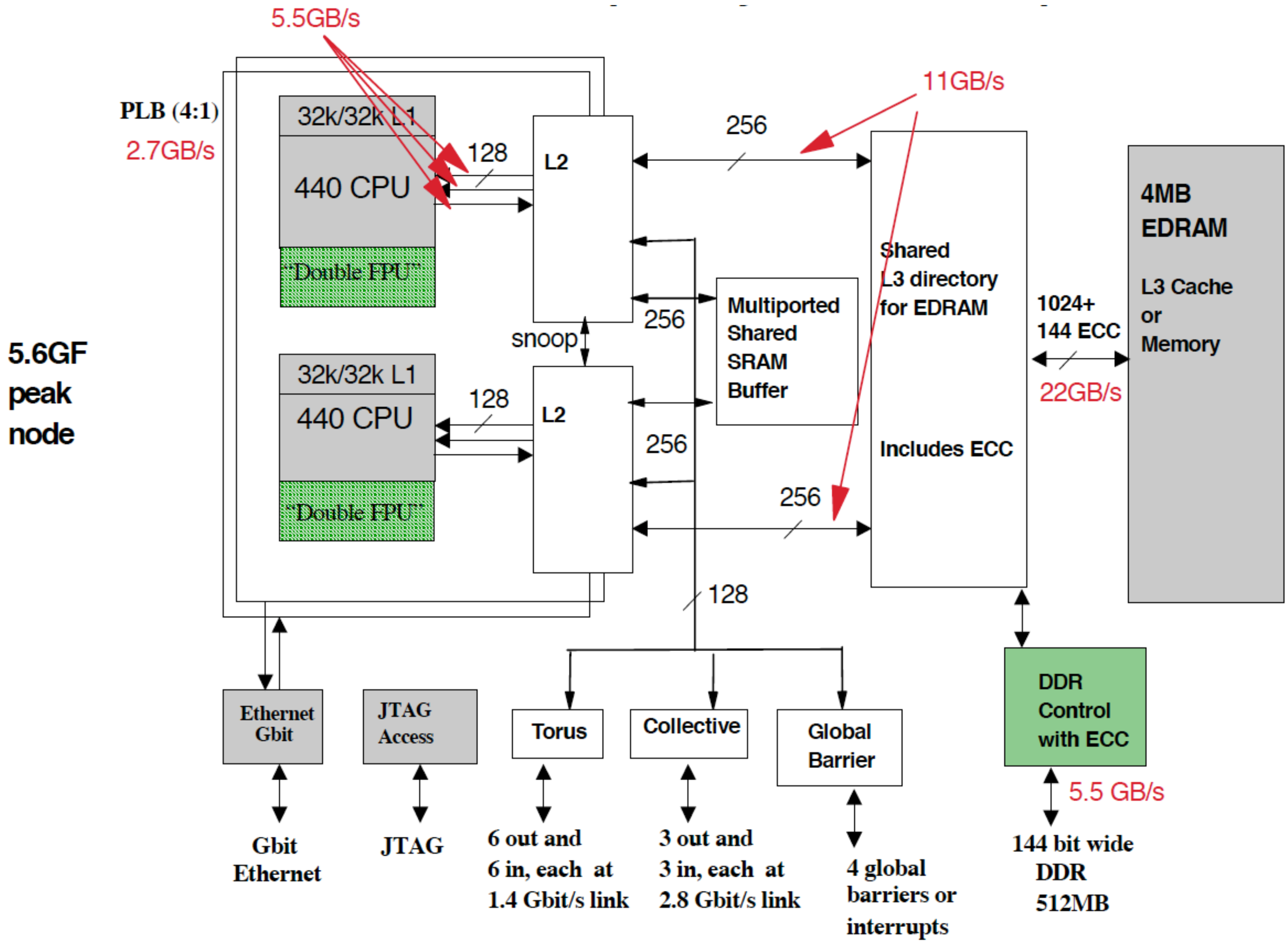


Dual FPU Architecture

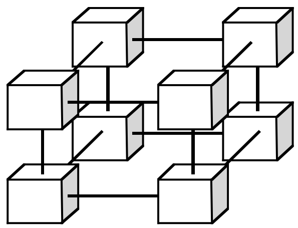


- Two 64 bit floating point units
- Designed with input from compiler and library developers
- SIMD instructions over both register files
  - FMA operations over double precision data
  - More general operations available with cross and replicated operands
    - *Useful for complex arithmetic, matrix multiply, FFT*
- Parallel (quadword) loads/stores
  - Fastest way to transfer data between processors and memory
  - Data needs to be 16-byte aligned
  - Load/store with swap order available
    - *Useful for matrix transpose*

BlueGene/L Compute System on a Chip ASIC

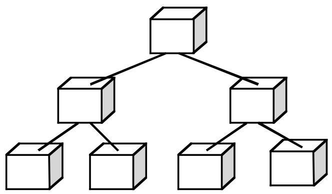


BlueGene/L - Five Independent Networks



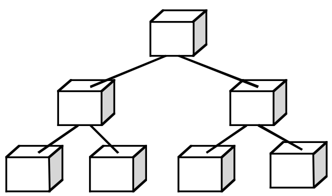
3 Dimensional Torus

- Point-to-point



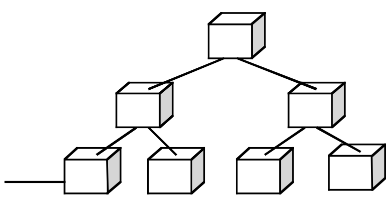
Global Tree

- Global Operations



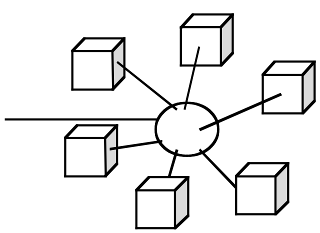
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

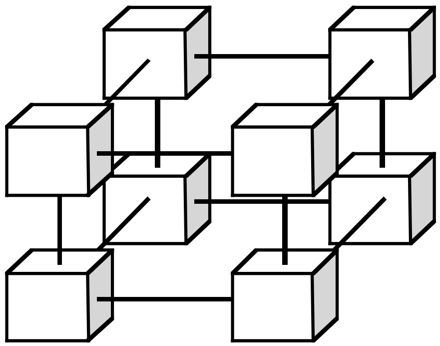
- File I/O and Host Interface

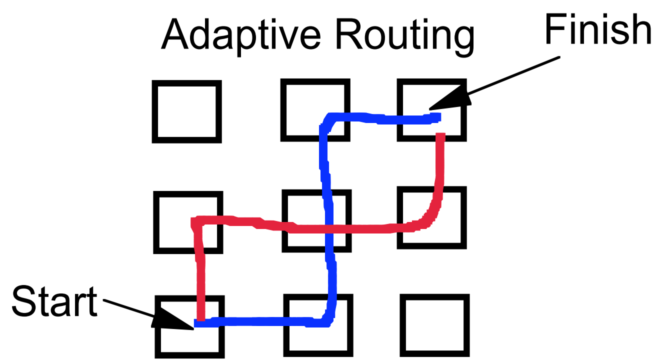


Control Network

- Boot, Monitoring and Diagnostics

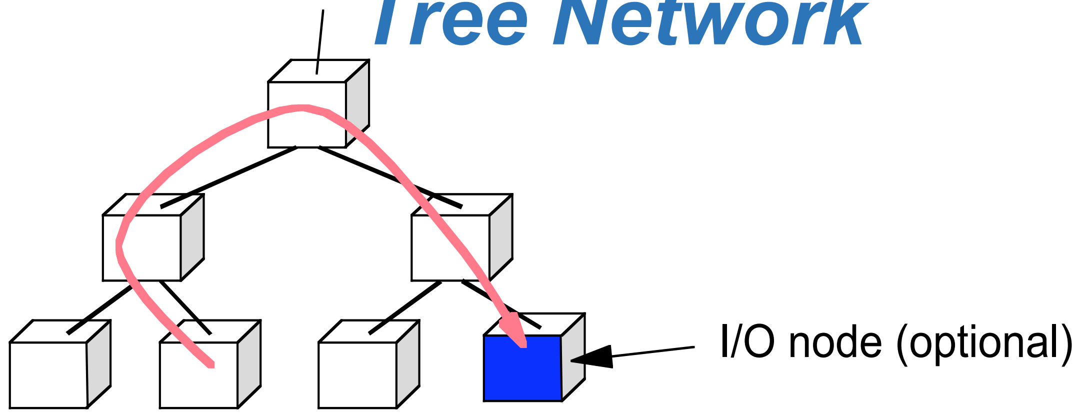
3-D Torus Network





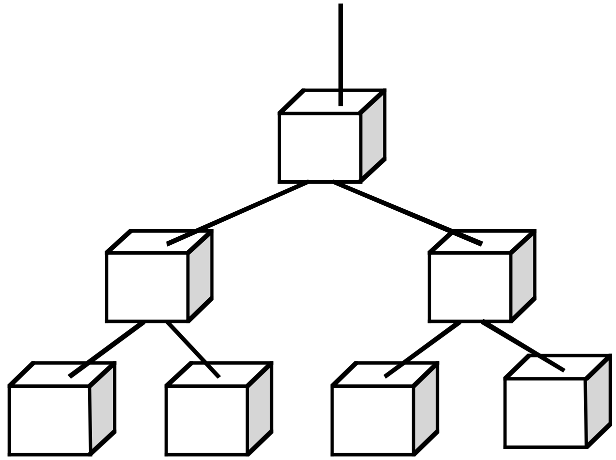
- 32x32x64 connectivity
- Backbone for one-to-one and one-to-some communications
- 1.4 Gb/s bi-directional bandwidth in all 6 directions (Total 2.1 GB/s/node)
- 64k * 6 * 1.4Gb/s = 68 TB/s total torus bandwidth
- 4 * 32 *32 * 1.4Gb/s = 5.6 Tb/s Bisectional Bandwidth
- Worst case hardware latency through node ~ 69nsec
- Virtual cut-through routing with multipacket buffering on collision
  - Minimal
  - Adaptive
  - Deadlock Free
- Class Routing Capability (Deadlock-free Hardware Multicast)
  - Packets can be deposited along route to specified destination.
  - Allows for efficient one to many in some instances
- Active messages allows for fast transposes as required in FFTs.
- Independent on-chip network interfaces enable concurrent access.

Tree Network



- High Bandwidth one-to-all
  - 2.8Gb/s to all 64k nodes
  - 68TB/s aggregate bandwidth
- Arithmetic operations implemented in tree
  - Integer/ Floating Point Maximum/Minimum
  - Integer addition/subtract, bitwise logical operations
- Latency of tree less than 2.5usec to top, additional 2.5usec to broadcast to all
- Global sum over 64k in less than 2.5 usec (to top of tree)
- Used for disk/host funnel in/out of I/O nodes.
- Minimal impact on cabling
- Partitioned with Torus boundaries
- Flexible local routing table
- Used as Point-to-point for File I/O and Host communications

Fast Barrier Network

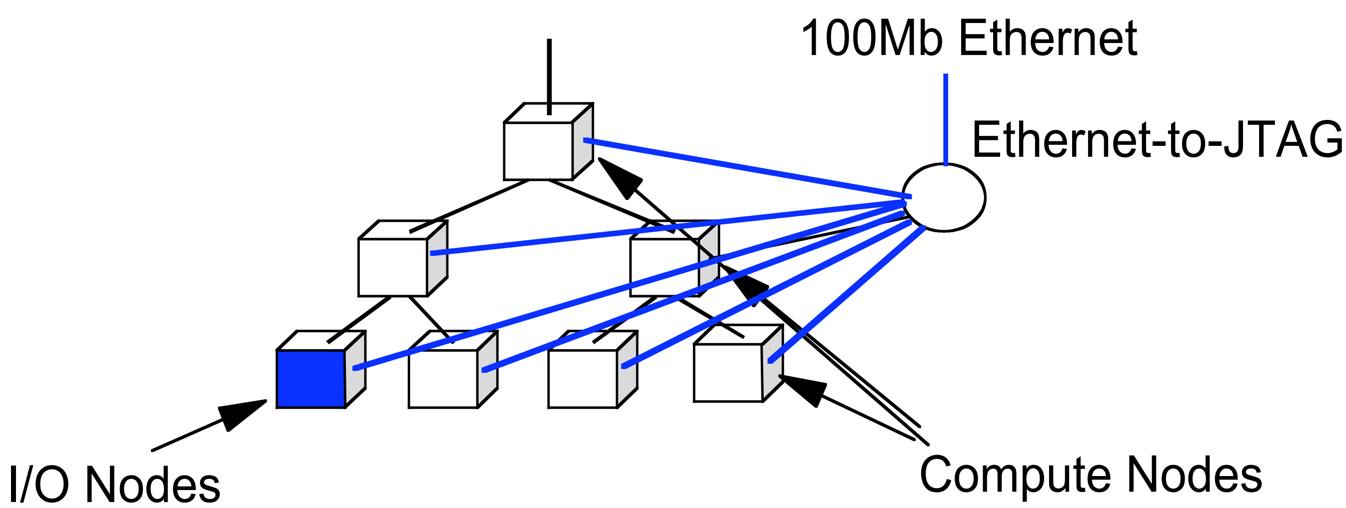


- Four Independent Barrier or Interrupt Channels
  - Independently Configurable as "or" or "and"
- Asynchronous Propagation
  - Halt operation quickly (current estimate is 1.3usec worst case round trip)
    > 3/4 of this delay is time-of-flight.
- Sticky bit operation
  - Allows global barriers with a single channel.
- User Space Accessible
  - System selectable
- Partitions along same boundaries as Tree, and Torus
  - Each user partition contains it's own set of barrier/ interrupt signals

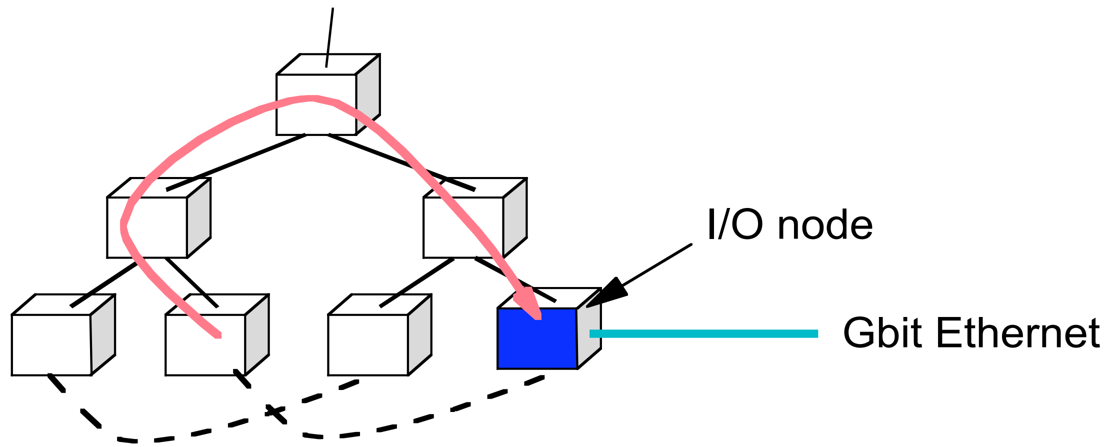
Control Network

JTAG interface to 100Mb Ethernet

- direct access to all nodes.
- boot, system debug availability.
- runtime noninvasive RAS support.
- non-invasive access to performance counters
- Direct access to shared SRAM in every node



Ethernet Disk/Host I/O Network



Gb Ethernet on all I/O nodes

- Gbit Ethernet Integrated in all node ASICs but only used on I/O nodes.
- Funnel via global tree.
- I/O nodes use same ASIC but are dedicated to I/O Tasks.
- I/O nodes can utilize larger memory.

Dedicated DMA controller for transfer to/from Memory

Configurable ratio of Compute to I/O nodes

- I/O nodes are leaves on the tree network

The Blue Gene Family of Computers

- Puts processors + memory + network interfaces on same chip.
- Achieves good compute-communications balance.

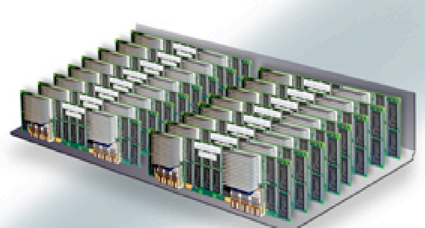
System
64 Racks, 64x32x32

180/360 TF/s
32 TB

Rack
32 Node Cards

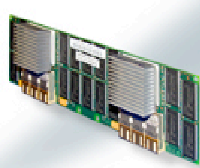
2.8/5.6 TF/s
512 GB

Node Card
(32 chips 4x4x2)
16 compute, 0-2 IO cards



90/180 GF/s
16 GB

Compute Card
2 chips, 1x2x1



5.6/11.2 GF/s
1.0 GB

Chip
2 processors

2.8/5.6 GF/s
4 MB

- Reaches high packaging density.
- Low system power requirements.
- Low cost per flops.

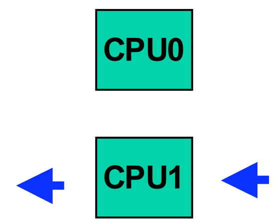
Record 280TF Linpack benchmark on 64K node BG/L at LLNL

Programming Environment

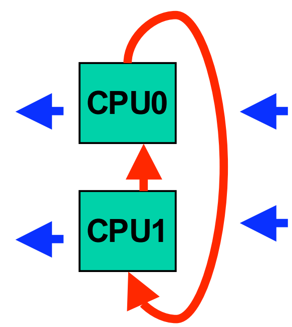
- Fortran, C, C++ with MPI
- Linux: User accesses system through Front End nodes for compilation, job submission, debugging
- Compute Node OS: very small, selected services, I/O forwarding
- No OpenMP, no Threads
- Space sharing - one parallel job (user) per partition of machine, one process per processor of compute node
- Single executable image is replicated on each node
- Virtual memory limited to physical memory
- Libraries are statically linked

Applications Developer's View of BlueGene

- Two CPU cores per node at 700 MHz
  - Each CPU can do 2 Float multiply-adds per cycle
- Mode 1 (Co-processor mode - CO)
  - CPU0 does all the computations (512MB memory)
  - CPU1 does the communications
  - Communications overlap with computation
  - Peak compute performance is 5.6/2 = 2.8 GFlops



- Mode 2 (Virtual node mode - VN)
  - CPU0, CPU1 independent "virtual tasks" (256MB each)
  - Each does own computation and communication
  - The two CPU's talk via memory buffers
  - Computation and communication cannot overlap
  - Peak compute performance is 5.6 GFlops



- 3-D torus network with virtual cut-through routing
  - (point to point: MPI_ISEND, MPI_IRECV)
- Global combine/broadcast tree network
  - (collectives: MPI_GATHER, MPI_SCATTER)

19 BlueGene Systems on 11/05 TOP500 List

Rank	Site	Country	Processors	RMax	RPeak
1	DOE/NNSA/LLNL	United States	131,072	280,600	367,000
2	IBM Thomas J. Watson Research Center	United States	40,960	91,290	114,688
9	ASTRON/University Groningen	Netherlands	12,288	27,450	34,406
12	Computational Biology Research Center, AIST	Japan	8,192	18,200	22,938
13	Ecole Polytechnique Federale de Lausanne	Switzerland	8,192	18,200	22,938
22	IBM - Rochester	United States	8,192	11,680	16,384
29	IBM - Almaden Research Center	United States	4,096	9,360	11,469
30	IBM - Deep Computing Capacity on Demand Cente	United States	4,096	9,360	11,469
31	IBM Research	Switzerland	4,096	9,360	11,469
32	IBM Thomas J. Watson Research Center	United States	4,096	9,360	11,469
73	Argonne National Laboratory	United States	2,048	4,713	5,734
74	Boston University	United States	2,048	4,713	5,734
75	Forschungszentrum Juelich (FZJ)	Germany	2,048	4,713	5,734
76	MIT	United States	2,048	4,713	5,734
77	NCAR (National Center for Atmospheric Research)	United States	2,048	4,713	5,734
78	NIWS Co, Ltd	Japan	2,048	4,713	5,734
79	Princeton University	United States	2,048	4,713	5,734
80	UCSD/San Diego Supercomputer Center	United States	2,048	4,713	5,734
81	University of Edinburgh	United Kingdom	2,048	4,713	5,734

In the first 6 months applications run on all 1024 nodes

Application	Institution	Domain	Description	Comments
Flash	ANL/UC	Astrophysics	Hydro (PPM) + Nuclear burning	Scaling tests to 16k processors
Nek5	ANL	General CFD	N-S using spectral elements	Good scaling to 2048 processors
QMC	ANL	Nuclear Physics	Nuclear binding energy using Monte Carlo	Significant science results. Good scaling
pNeo	ANL/UC	Neuroscience	Hodgkin/Huxley Model for neuron firing	Run to 2048. Optimizing comms.
DL_POLY	Daresbury Laboratory	Nano-Chemistry	Molecular dynamics with provisions for periodic slabs and solids	Good scaling to 2048 processors
Petsc FUN3d	ANL/NASA	General CFD	Unstructured Navier-Stokes solver	Good scaling to 2048 processors
POP	LANL	Oceanography	Primitive equations on sphere – hydro-static, Boussinesq	Run to 2048 procs.
Nimrod	U Wisconsin	Fusion	Non-ideal MHD (finite element) w/ rotation, complex boundaries	Near 1 TF/s on full machine
GTC	PPPL	Plasma Physics	Gyrokinetic toroidal particle-in-cell	Scaled to full system with good success
LSMS	PSC	Electronic Structure	Interactions between electrons and atoms in magnetic materials	Perfect weak scaling
FDTD	IBM Almaden	Nanophotonics	Finite difference time domain	Very good scaling
RXMD	USC	Molecular Dynamics	Dynamics of chemically reacting mixtures	Runs well on 2048 processors
EDC-DFT	USC	Elect. Structure	Quantum mechanics based molecular dynamics	Very good scaling
GibTigs	BU	Bioinformatics	Gibbs Sampling Monte Carlo Markov Chains	Very promising big runs

The Blue Gene/L Consortium

formed by Argonne and IBM, April 2004

- Focuses interest in the Blue Gene series
  - Exploiting its potential for computational science
- Creates a framework for cooperation
  - Developing applications, tools and systems software
  - Sharing support of systems (not a fully supported IBM product)
  - Exchanging innovations and novel solutions
- Supports upcoming HPC needs
  - Training students and develop next generation user community
  - Providing functional requirements for next generation systems

Working Groups

- **Applications**
- **System Software**
- **Operations**
- **Architecture**
- **Outreach**

http://www.mcs.anl.gov/bgconsortium/

ARGONNE NATIONAL LABORATORY

Blue Gene/L Consortium Members (55)

DOE Laboratories

- Ames National Laboratory/Iowa State U.
- Argonne National Laboratory
- Brookhaven National Laboratory
- Fermi National Laboratory
- Jefferson Laboratory
- Lawrence Berkeley National Laboratory
- Lawrence Livermore National Laboratory
- Oak Ridge National Laboratory
- Pacific Northwest National Laboratory
- Princeton Plasma Physics Laboratory

Universities

- Boston University
- California Institute of Technology
- Columbia University
- DePaul University
- Harvard University
- Illinois Institute of Technology
- Indiana University
- Louisiana State University
- Massachusetts Institute of Technology
- National Center for Atmospheric Research
- New York University/Courant Institute
- Northern Illinois University
- Northwestern University
- Ohio State University
- Pennsylvania State University
- Pittsburgh Supercomputing Center
- Princeton University

Universities (continued)

- Purdue University
- Rutgers University
- Stony Brook University (SUNY)
- Texas A&M University
- University of California
  Irvine, San Francisco, San Diego/SDSC
- University of Chicago
- University of Colorado
- University of Delaware
- University of Illinois – Urbana Champaign
- University of Minnesota
- University of North Carolina
- University of Southern California/ISI
- University of Texas at Austin – TACC
- University of Utah
- University of Wisconsin

Industry

- Engineered Intelligence Corporation
- IBM

International

- ASTRON/LOFAR, The Netherlands
- Centre of Excellence for Applied Research and Training, UAE
- Ecole Polytechnique Fédérale de Lausanne, Switzerland
- National University of Ireland
- Trinity College, Ireland
- John von Neumann Institute, Germany
- NIWS Co., Ltd., Japan
- University of Edinburgh, EPCC Scotland